世界知的所有権機関
国際事務局

PCT

特許協力条約に基づいて公開された国際出願

(51) 国際特許分類6
G06F 17/30

A1

(11) 国際公開番号 WO97/09683

(43) 国際公開日 1997年3月13日(13.03.97)

(21) 国際出願番号 PCT/JP95/01746

(22) 国際出願日 1995年9月1日(01.09.95)

(71) 出願人（米国を除くすべての指定国について）
株式会社 日立製作所(HITACHI, LTD.)[JP/JP]
〒101 東京都千代田区神田駿河台四丁目6番地 Tokyo, (JP)
(72) 発明者；および
(75) 発明者／出願人（米国についてのみ）
菊池英明(KIKUCHI, Hideaki)[JP/JP]
〒185 東京都国分寺市東恋ヶ窪3-1-3
日立第2協心寮 Tokyo, (JP)
畑岡信夫(HATAOKA, Nobuo)[JP/JP]
〒220-01 神奈川県津久井郡城山町町屋4丁目15の2
Kanagawa, (JP)
在~~原~~俊之(ARITSUKA, Toshiyuki)[JP/JP]
〒189 東京都東村山市多摩湖町4-23-13 Tokyo, (JP)
(74) 代理人
弁理士 小川勝男(OGAWA, Katuo)
〒100 東京都千代田区丸の内一丁目5番1号
株式会社 日立製作所内 Tokyo, (JP)

(81) 指定国 CN, JP, KR, US, 欧州特許 (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

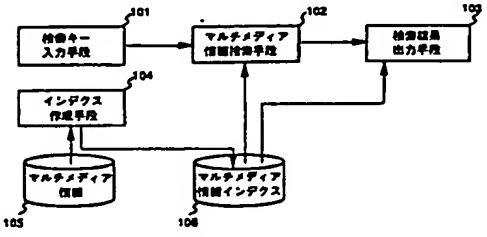
添付公開書類
国際調査報告書

(54)Title: AUTHORING SYSTEM FOR MULTIMEDIA INFORMATION INCLUDING SOUND INFORMATION

(54)発明の名称 音声情報を含むマルチメディア情報のオーサリング方式

(57) Abstract

An authoring system by which retrieval of a moving picture or sound information from video information including sound information is facilitated using a portable information terminal such as a PDA (Personal Digital Assistant) notebook computer, or using a multimedia terminal such as a personal computer or a workstation. The authoring system is provided with at least a retrieval key-inputting means through which a retrieval key such as a key word or an attribute value is inputted, retrieval result outputting means which outputs the retrieved sound information or moving picture, multimedia information retrieving means which retrieves multimedia information including sound information and moving picture information, and index generating means which generates indexes representing the correspondences between sound information and the moving picture information with respect to multimedia information including sound information. A desired moving picture or sound information can be readily retrieved from other corresponding information.



101: retrieval-key inputting means
102: multimedia information retrieving means
103: retrieval result outputting means
104: index generating means
105: multimedia information
106: multimedia information index

SON-0522

A

(19)日本国特許庁（ＪＰ）　**再公表特許（Ａ１）**　(11)国際公開番号

WO97／09683

発行日　平成10年（1998）10月20日　(43)国際公開日　平成9年（1997）3月13日

(51)Int.Cl.[6]　識別記号　ＦＩ

Ｇ０６Ｆ　17/30

審査請求　未請求　予備審査請求　有　（全 27 頁）

出願番号　特願平9－511051

(21)国際出願番号　PCT／JP95／01746

(22)国際出願日　平成7年（1995）9月1日

(81)指定国　ＥＰ（ＡＴ，ＢＥ，ＣＨ，ＤＥ，ＤＫ，ＥＳ，ＦＲ，ＧＢ，ＧＲ，ＩＥ，ＩＴ，ＬＵ，ＭＣ，ＮＬ，ＰＴ，ＳＥ），ＣＮ，ＪＰ，ＫＲ，ＵＳ

(71)出願人　株式会社日立製作所
東京都千代田区神田駿河台4丁目6番地

(72)発明者　菊池　英明
東京都国分寺市東恋ヶ窪3－1－3　日立第2協心寮

(72)発明者　畑岡　信夫
神奈川県津久井郡城山町町屋4丁目15の2

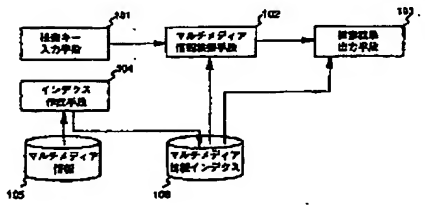
(72)発明者　在塚　俊之
東京都東村山市多摩湖町4－23－13

(74)代理人　弁理士　小川　勝男

(54)［発明の名称］音声情報を含むマルチメディア情報のオーサリング方式

(57)【要約】
本発明は、ＰＤＡ(Personal Digital Assistant)ノートパソコンなどの携帯情報端末や、パーソナルコンピュータ、ワークステーションなどのマルチメディア端末において、音響情報を含む映像に対して、動画像や音声の検索を容易にするオーサリング方式を提供する。少なくとも、キーワードや属性値などの検索キーを入力する検索キー入力手段と、音響情報あるいは動画像を検索結果として出力する検索結果出力手段と、音響情報と動画像情報を含むマルチメディア情報を検索するマルチメディア情報検索手段とを有し、音響情報を含むマルチメディア情報について音響情報と動画像との対応関係を示すインデクスを作成するインデクス作成手段を備えることにより、欲しい動画あるいは欲しい音響情報を、対応する他の情報から容易に検索することを可能にした。

第１図



[特許請求の範囲]

1．音響情報と動画像情報を含むマルチメディア情報を記憶する手段（１０５）と、

上記マルチメディア情報を読みだして音響情報と動画像との対応関係を示すインデクスを作成するインデクス作成手段（１０４）と、

上記インデクスを記憶する手段（１０６）と、

欲しい動画像あるいは欲しい音響情報に関する検索情報を入力するための検索キー入力手段（１０１）と、

上記インデクスを参照して上記検索情報に対応する動画像又は音響情報を検索するマルチメディア情報検索手段（１０２）と、

上記検索結果を出力する検索結果出力手段（１０３）と

からなるマルチメディア情報のオーサリング方式。

2．前記インデクス作成手段は、

上記マルチメディア情報に含まれる音響情報の音声区間を検出する音声区間検出手段（２０１）と、

該音声区間をもとに音声インデクスを作成する音声インデクス作成手段（２０２）とを有する請求の範囲第１項に記載のマルチメディア情報のオーサリング方式。

3．上記検索結果出力手段は、インデクス表示手段及びディスプレイを有し、上記検索結果及び上記インデクスを表示する請求の範囲第１項に記載のマルチメディア情報のオーサリング方式。

4．上記検索結果出力手段は、インデクス表示手段及びディスプレイを有し、上記検索結果及び上記インデクスを表示し、

上記ディスプレイ上に表示されたインデクスを用いて指定された音声区間を検索情報として指定する請求の範囲第１項に記載のマルチメディ

ア情報のオーサリング方式。

5．上記検索情報を任意の動画像とする請求の範囲第１項に記載のマルチメディア情報のオーサリング方式。

6．前記インデクス作成手段は、

マルチメディア情報に含まれる音響情報の音声区間を検出する音声区間検出手段（２０１）と、

該音声区間検出手段により検出した音声区間の音声について話者を識別し、全音声区間について該話者を識別する話者識別手段（８０１，８０２）と、

該話者と前記音声区間をもとに音声インデクスを作成する音声インデクス作成手段（２０２）と、

からなる

請求の範囲第１項のマルチメディア情報のオーサリング方式。

7．上記検索情報を人物名として、該人物の音声区間の音声あるいは該音声が対応する人物画像を検索することを特徴とする、請求範囲第１項のマルチメディア情報のオーサリング方式。

8．前記マルチメディア情報検索手段（１０２）は、

動画像内の人物画像から口唇の動きを検出し、口唇の動きに対応する音素を識別する口唇認識手段（１５０１）と、

動画像内の音声情報を音素標準パタンにもとづき認識する音声認識手段（１５０６）と、

該口唇認識手段が出力する音素識別結果と、音声認識手段が出力する音声認識結果を比較照合する画像音声照合手段（１５０２）と、

該画像音声照合手段において、前記音素識別結果が一致すると判定された音声区間の動画像を抽出するシーン抽出手段（１５０３）と、

を有し、

音声区間の音声に対応する人物画像または、人物画像に対応する音声区間の音声を得る請求範囲第１項のマルチメディア情報のオーサリング方式。

9．上記検索情報を動画像内の人物画像とし、該人物画像の音声区間の音声あるいは動画像を検索することを特徴とする請求範囲第１項のマルチメディア情報のオーサリング方式。

１０．音声送信要求プロトコルを発信する音声送信要求手段（６０２）と、

音声に対して検索を行なう音声検索手段（６０６）と、

動画像送信要求プロトコルを発信する動画像送信要求手段（６０７）と、

を備えたマルチメディア情報表示クライアント（以下、クライアント）と、

音響送信要求プロトコルを受信し、該プロトコルにおいて指定されたマルチメディア情報を取得する情報取得手段（６０３）と、

マルチメディア情報から音声を抽出する音声抽出手段（６０４）と、

音声を送信する音声送信手段（６０５）と、

動画像を送信する動画像送信手段（６０９）と、

を備えたマルチメディア情報検索サーバ（以下、サーバ）と、

を有するマルチメディア情報検索クライアントサーバシステムにおいて、

上記サーバは、動画像送信要求プロトコルを受信した後、該プロトコルにおいて指定された区間の動画像を抽出するシーン抽出手段（６０８）を有し、

マルチメディア情報のうち、全ての情報を通信することなく所望の区間の情報のみを通信するマルチメディア情報検索クライアントサーバシ

ステム。

[発明の詳細な説明]

音声情報を含むマルチメディア情報のオーサリング方式

技術分野

本発明は、ＰＤＡ（Personal Digial Assistant）ノートパソコンなどの携帯情報端末や、パーソナルコンピュータ、ワークステーションなどのマルチメディア端末において、音響情報を含む映像に対して、話者別の映像を容易に抽出することを可能にしたオーサリング方式を提供する。

背景技術

従来の映像オーサリング方式において、映像から人物別シーンを抽出する場合、画像フレームから人物画像を抽出するために、人物画像と音声との対応がとれず、抽出したシーンの区間は必ずしもその人物の音声区間と一致しないという問題があった。これに対して、あらかじめ人物別に画像特徴量と音声特徴量を保有し、それぞれの特徴量から人物画像識別、音声話者識別を行ない、同一人物の人物画像と音声を対応づける手法が考えられるが、現実的には人物別の画像特徴量と音声特徴量を保有することは不可能であり、実現性は低い。

従来技術では、人物画像と、それに対応する音声区間を含むシーンを自動的に抽出することは困難である。

本発明の目的は、マウスによる画像からの人物指定や、キーボードによる人物名入力により、該当する人物の画像出現区間と発話音声区間を含むシーンを、自動的に抽出できるシステムを提供することである。

発明の開示

上記の課題を解決するために、本発明のマルチメディア情報オーサリング方式では、少なくとも、キーワードや属性値などの検索キーを入力する検索キー入力手段と、音響情報あるいは動画像を検索結果として出力する検索結果出力手段と、音響情報と動画像情報を含むマルチメディア情報を検索するマルチメディア情報検索手段と、を有し、音響情報を含むマルチメディア情報について音響情報と動画像との対応関係を示すインデクスを作成するインデクス作成手段を備えることにより、欲しい動画像あるいは欲しい音響情報を、対応する他の情報から容易

に検索することを可能にした。

前記インデクス作成手段は、マルチメディア情報に含まれる音響情報の音声区間を検出する音声区間検出手段と、該音声区間をもとに音声インデクスを作成する音声インデクス作成手段と、を有し、音声区間の音声に対応する動画像または、動画像に対応する音声区間の音声を容易に得ることを可能にした。

前記マルチメディア情報のインデクスをディスプレイに表示するインデクス表示手段を有することにより、マルチメディア情報のオーサリングを視覚的に行なうことを可能にした

前記インデクス表示手段によりディスプレイ上に表示されたインデクスに対して、音声区間を指定することにより、音声区間の音声あるいは動画像を検索する。動画像の任意の画像を指定することにより、前記インデクス作成手段により作成されたインデクスを用いて、指定画像に対応する音声区間の手声あるいは動画像を検索する。

マウスなどの位置入力手段を用いて、所望のマルチメディア情報の範囲を指定し、別ウインド内の任意の位置を前記位置入力手段により指定することにより、前記マルチメディア情報への参照情報を該位置に加え

ことを可能にする、ハイパーリング型マルチメディア情報のオーサリング方式を構成することもできる。

前記インデクス作成手段は、マルチメディア情報に含まれる音響情報の音声区間を検出する音声区間検出手段と、該音声区間検出手段により検出した音声区間の音声について話者を識別し、全音声区間について該話者を識別する話者識別手段と、該話者と前記者声区間をもとに音声インデクスを作成する音声テンデクス作成手段と、を有することにより、同一話者の全音声区間の音声に対応する動画像または、動画像に対応する同一話者の全音声区間の音声を容易に得ることを可能にした。

キーボードなどの文字入力手段を用いて、人物名を指定することにより、該人物の音声区間の音声あるいは動画像を検索する。

前記マルチメディア情報検索手段は、動画像内の人物画像から口唇の動きを検

出し、口唇の動きに対応する音素を識別する口唇認識手段と、動画像内の音声情報を音素標準パタンにもとづき認識する音声認識手段と、該口唇認識手段が出力する音素識別結果と、音声認識手段が出力する音声認識結果を比較照合する画像音声照合手段と、該画像音声照合手段において、前記音素識別結果が一致すると判定された音声区間の動画像を抽出するシーン抽出手段と、を有することにより、音声区間の音声に対応する人物画像または、人物画像に対応する音声区間の音声を容易に得ることができる。

前記マルチメディア情報検索手段は、マウスなどの位置入力手段によって入力された位置に応じて、動画像内の該位置に存在する人物画像を抽出する人物画像抽出手段を有することにより、前記位置入力手段を用いて動画像内の人物画像を指定し、自動的に該人物の音声区間の音声あるいは動画像を検索することができる。

また、本発明のマルチメディア情報検索クライアントサーバシステム

は、音声送信要求プロトコルを発信する音声送信要求手段と、音声に対して検索を行なう音声検索手段と、動画像送信要求プロトコルを発信する動画像送信要求手段と、を備えたマルチメディア情報表示クライアント（以下、クライアント）と、音響送信要求プロトコルを受信し、該プロトコルにおいて指定されたマルチメディア情報を取得する情報取得手段と、マルチメディア情報から音声を抽出する音声抽出手段と、音声を送信する音声送信手段と、動画像を送信する動画像送信手段と、を備えたマルチメディア情報検索サーバ（以下、サーバ）と、を有し、さらにサーバは、動画像送信要求プロトコルを受信した後、該プロトコルにおいて指定された区間の動画像を抽出するシーン抽出手段を有し、マルチメディア情報のうち、全ての情報を通信することなく所望の区間の情報のみを通信することを可能にした。

図面の簡単な説明

第１図はマルチメディア情報オーサリング方式の全体構成図であり、第２図はインデクス作成手段の構成例であり、第３図はインデクス作成手段の他の構成例であり、第４図はマルチメディア情報検索手段の構成例であり、第５図はマルチ

メディア情報検索手段の他の構成例であり、第６図は検索結果出力手段の構成例であり、第７図は本発明の画面表示例であり、第８図は本発明の画面表示の他の例であり、第９図はマルチメディア情報検索クライアントサーバシステムの構成例であり、第１０図はマルチメディア情報検索クライアントサーバシステムの他の構成例であり、第１１図は本発明の画面表示例である。

発明を実施するための最良の形態

以下、図を用いて実施例を詳細に説明する。なお、以下、マルチメデ

ィア情報は少なくとも音声および動画像を含む情報とする。また、ここでは、特にマルチメディア端末として、マルチメディア情報のブラウズと編集の機能を持つ携帯情報端末を想定して説明を行う。ただし、本発明は該携帯情報端末に限らず、パーソナルコンピュータやワークステーションなどのマルチメディア端末や、編集機能を持つ家庭用、英会話学習用ビデオデッキ、ＴＶ電話留守録ビデオなどの映像蓄積機能を持つマルチメディア情報機器一般への応用が可能である。

第１図は、本発明のマルチメディア情報オーサリング方式のブロック構成図である。

第１図において、検索キー入力手段１０１は、利用者が編集する対象を検索するために、検索のキーとなるキーワードや位置などを入力する手段である。マルチメディア情報検索手段１０２は、マルチメディア情報に対して任意の区間の音声あるいは動画像を検索する手段である。検索結果出力手段１０３、マルチメディア情報検索手段１０２の検索結果を、利用者に提示するために出力する手段である。インデクス作成手段１０４は、マルチメディア情報について音響情報と動画像との対応関係を示すインデクスを作成する手段である。

具体的には、マルチメディア情報１０５に含まれる音声について、音声が存在する音声区間や音声と対応する話者名で区間わけをする。また、動画像について、例えば、画像内の人物毎にそれぞれの人物に対応した区間分けを行なう、などの任意の規則に基づいた区間分けによる動画表示区間を用いる。なお、インデクス作成手段１０４も利用者により実施される場合とマルチメディア情報蓄積後の任意の時期に自動的に実施される場合が考えられる。以下では、利用者により実

施される場合を想定する。

利用者は、まず、検索キー入力手段１０１を用いて編集する対象を検

索するために検索のキーを入力する。ここで、検索のキーとしては、文字列や、静止画像内の任意の部分画像、区間などが考えられる。検索キー入力手段１０１は、これらの検索キーの全てについて単独あるいは複合入力可能とする。次に、マルチメディア情報検索手段１０２は、検索キー入力手段１０１により入力された検索キーを用いて、マルチメディア情報インデクス１０６に対して、検索キーと合致するインデクスを持つ特定の区間のマルチメディア情報を検索する。さらに、検索結果出力手段１０３は、マルチメディア情報のインデクスをディスプレイに表示したり、マルチメディア情報検索手段１０２により検索された音声、あるいは動画像を出力する。具体的には、音声の場合にはスピーカ、ヘッドフォンなどから音声出力し、動画像の場合には、ディスプレイなどへの表示を行なう。

例えば、検索キー入力手段１０１によって、マルチメディア情報に含まれる音声のうち、特定の区間を示す検索キーが入力された場合、マルチメディア情報検索手段１０２では、あらかじめインデクス作成手段１０４によって話者別の音声区間に基づいて作成されたマルチメディア情報インデクス１０６を用いて、音声あるいは音声に対応する動画像を検索する。検索された音声あるいは動画像は、検索結果出力手段１０３により、音声出力あるいは動画像表示が行なわれる。

第２図に、本発明のインデクス作成手段１０４の構成例を示す。第２図において、マルチメディア情報インデクス１０６として音声インデクス２０４を作成する。したがって、インデクス作成手段１０４は、音声区間検出手段２０１と音声インデクス作成手段２０２とから構成されている。

音声区間検出手段２０１は、蓄積されたマルチメディア情報２０３に含まれる音響情報に対して、人間の音声区間を検出する手段である。音

響情報における音声区間の検出を行なう方法として、例えば、一定のしきい値以上の値の短時間パワーが一定時間以上継続したか否かが用いられる方法がある（“デジタル音声処理”、東海大学出版会、ｐｐ１５３「8.2　音声区間の検出

」参照）．音声インデクス作成手段２０２は、音声区間検出手段２０１により検出した音声区間の情報をもとにインデクスを作成する．ここで、音声インデスク作成手段２０２により作成されるインデクスは、例えば、検出された各音声区間の始端、終端の時刻や、音声区間長などが挙げられる．

このように音声区間に基づいたインデクスを作成することにより、音声区間の音声に対応する動画像またはその逆として、動画像に対応する音声区間の音声を容易に得ることができるようになる．

第３図は、本発明のインデクス作成手段１０４の他の構成例である．第３図において、話者識別手段８０１は、音声に対して、特定の話者の音声標準パタンとの照合を行ない、音声が指定された話者の音声であるかを識別する手段である．話者識別の方法として、例えば、音声波から特徴抽出をしたのち、あらかじめ蓄えられている各登録話者の標準パタンとの距離あるいは類似度を調べ、その度合いにより識別の判定を行なう方法がある（"ディジタル音声処理"，東海大学出版会、ｐｐ１９６「９．３　話者認識系の構成」参照）．

第３図において、まず、音声区間検出手段２０１により、蓄積されたマルチメディア情報２０３の音響情報に対して人間の音声区間を検出する．さらに、検出した音声区間の音声について、話者識別手段８０１により、音声標準パタン８０２に基づいた話者識別を行なう．話者識別を行なった結果、各音声区間の音声に対して、該当する話者名を得る．従って、音声インデクス作成手段２０２により、音声区間と話者名を関連付けて、マルチメディア情報インデクスとしてインデクス２０４を作成

する．ここで、音声インデクス作成手段２０２により作成されるインデクスは、例えば、検出された各音声区間の始端、終端の時刻や、音声区間長と話者名などが挙げられる．

このように音声区間の話者名に基づいたインデクスを作成することにより、音声区間の音声に対応する動画像またはその逆として、動画像に対応する音声区間の音声を容易に得ることができるようになる．

第４図は、本発明のマルチメディア情報検索手段１０２の構成例を示す図であ

る．第４図において、口唇認識手段１５０１は、入力画像において抽出された人物顔画像から唇の動きを認識し、唇の動きに対応する音素を出力する手段である．唇の動きから音素を認識する方法として、例えば、まず画像処理による２次元形状抽出を行ない、そのデータに対してニューラルネットを用いて音素識別を行なう方法がある（"ノンバーバルインターフェース"，オーム社、ｐｐ１４９「口話の認識」参照）．音声認識手段１５０６は、音声情報について音声認識を行なう手段である．なお、入力音声の音声認識を行なう方法として、例えば、入力音声を小区間ごとに音素の標準パタンと比較して距離を求め、距離の近い音素を音素認識結果として出力し、さらに音素系列を単語音声辞書と比較する手段がある（前出"ディジタル音声処理"，東海大学出版、ｐｐ１６７「８．６　音素を単位とする単語音声認識」参照）．画像音声照合手段１５０２は、人物画像における唇の動きに対応する音素系列と、入力音声の照合を行なう手段である．シーン抽出手段１５０３は、指定された区間の映像を切り出す手段である．

第４図において、まず、口唇認識手段１５０１において、口形や口面積などの特徴量の標準パターン１５０４との照合により唇の動きを認識する．なお、口唇認識の結果としては、音素系列を出力することによる．次に、音声認識手段１５０６において、音素区間内の音声のスペクトル

と音素標準パタン辞書１５０７の各音素スペクトルとの類似度計算により音素系列を音声認識結果として出力する．ここで、画像音声照合手段１５０２において、口唇認識手段１５０１の出力結果である音素系列と、音声認識手段１５０６の出力結果である音素系列の比較照合を行なう．これにより、人物画像における唇の動きと前後の音声区間とを照合し対応付けることができる．最後に、シーン抽出手段１５０３において、人物画像に対応づけられた音声区間の映像を、全映像から抽出する．

以上の処理により、ペンなどの位置入力手段によって指定された人物の映像について、音声区間を含む映像区間を入力像から抽出することが可能となる．また、同一話者の全音声区間に対応する人物画像、または、人物画像に対応する同話者の全音声区間の音声を容易に得ることが可能となる．

第５図は、本発明のマルチメディア情報検索手段１０２の他の構成例を示す図である．第５図では、人物画像抽出手段１９０１を設け入力画像から自動的に人物の有無を検出し、人物の顔を検出する．入力画像から自動的に人物の有無の検出、さらに顔の検出を行なう方法として、例えば、複数の解像度で画像をサンプリングして得られるピラミッド画像を照合する方法などがある（"ディジタル信号処理ハンドブック"，電子情報通信学会刊、ｐｐ４０１「４．３．３　人物の認識」参照）．口唇認識手段１９０２は、入力画像において抽出された人物顔画像から唇の動きを認識し、唇の動きに対応する音素を出力する手段である．音声認識手段１９０７は、音声情報について音声認識を行なう手段である．画像音声照合手段１９０３は、人物画像における唇の動きに対応する音素系列と、入力音声の照合を行なう手段である．シーン抽出手段１９０４は、指定された区間の映像を切り出す手段である．

第５図において、まず位置入力手段を用いて入力された画面上の位置

座標をもとに、人物画像抽出手段１９０１において、入力画像内の入力位置座標付近の領域について人物画像の有無を検出し、さらに人物顔画像を抽出する．なお、入力画像内に一つの人物画像が検出された場合には、それを指定画像とし、入力画像内に複数の人物画像が検出された場合には、位置入力手段１０１により入力された座標点を含む、もしくは最も近い人物画像を指定画像とする．人物画像抽出手段１０２によって抽出された人物顔画像について、次に、口唇認識手段１９０２において、口形や口面積などの特徴量の標準パタン１９０５との照合により唇の動きを認識する．なお、口唇認識の結果としては、音素系列を出力することにする．次に、音声認識手段１９０７において、音素区間内の音声のスペクトルと音素標準パタン部　１９０８の各音素スペクトルとの類似度計算により音素系列を音声認識結果として出力する．

ここで、画像音声照合手段１９０３において、口唇認識手段１９０２の出力結果である音素系列と、音声認識手段１９０７の出力結果である音素系列の比較照合を行なう．これにより、人物画像における唇の動きと前後の音声区間とを照合し対応付けることができる．最後に、シーン抽出手段１９０４において、人物画

像に対応づけられた音声区間の映像を、全映像から抽出する．

以上の処理により、ペンなどの位置入力手段によって指定された人物の映像について、音声区間を含む映像区間を入力映像から抽出することが可能となる．

第６図は、本発明のマルチメディア情報オーサリング方式におけるインデクス表示を行なうブロック構成例を示す図である．第６図において、インデクス作成手段３０３は、第１図におけるインデクス作成手段１０４に対応する．インデクス表示手段３０１は、マルチメディア情報インデクスを視覚化してディスプレイに表示する手段である．

第６図において、まず、インデクス作成手段３０３により作成されたマルチメディア情報インデクス３０４について、インデクス表示手段３０１によって視覚化を行ない、ディスプレイ３０２に表示する．例えば、音声区間に基づいて作成されたインデクスについて、横軸に時刻をとった２次元座標系に、各音声区間の始端、終端の時刻や区間長を棒線による表示方法が考えられる．あるいは、話者別に区間分けされた音声のインデクスに関しては、さらに話者別に棒線をを配置して表現する方法が考えられる．

なお、具体的には、第１図の検索結果出力手段１０３は、インデクス表示手段３０１、ディスプレイ３０２から構成されている．

このようなインデクスの視覚化を行なうことにより、マルチメディア情報のオーサリングを視覚的に行なうことが可能になる．

第７図には、インデクスを視覚化した画面表示例を示す図である．第７図において、映像表示領域４０１は、ディスプレイ上の、動画像を表示する領域である．インデクス表示領域４０２は、ディスプレイ上の、マルチメディア情報インデクスを表示する領域である．音声インデクス表示領域４０３は、ディスプレイ上の、音声インデクスを表示する領域である．指定音声区間４０４は、利用者が音声あるいは動画像の出力を要求するために指定した音声区間を示す．指定画像４０５は利用者が音声あるいは動画像の出力を要求するために指定する画像を示す．

第７図において、まず、利用者はインデクス表示領域内の音声インデクス表示

領域４０３に表示された音声インデクスに対して、欲しい音声あるいは動画像に対応する音声区間を指定することにより、音声あるいは動画像を出力させることができる。また、利用者は、映像表示領域４０１内に表示されている動画像に対して、現在出力されている音声に対応する音声区間あるいは動画像を要求する場合、画像４０５を指定する

ことにより、要求した音声区間の音声あるいは動画像を出力させることができる。

他の表示例として第８図に、本発明のマルチメディア情報オーサリング方式を携帯端末に利用した際の画面表示例を示す。第８図において、携帯情報端末の画面上に、映像表示領域７０２と、文書表示領域７０３と、メニュー領域７０１を設けている。まず、第８図の左側の携帯情報端末上で、メニュー領域７０１内から「セリフ抽出」という項目を選択する。次に、映像表示領域７０２上で映像再生中に、セリフを抽出したい人物画像の位置を位置入力手段７０５によって指定する。ここまでの操作により、第１図において示したマルチメディア情報のオーサリング方式を用いて指定された人物画像に対応する音声区間を含むシーンを抽出する。第８図の右側の携帯情報端末上では、さらに、抽出したシーンをシンボル化したアイコン７０４を、マウスなどの位置入力手段を用いて画面上で動かし、文書表示領域７０３内の任意の位置にアイコン７０４をおくことにより、文書表示領域７０３上の文書と、抽出した映像を関連付ける操作を示した。

第９図に、本発明のマルチメディア情報オーサリング方式を利用したマルチメディア情報検索クライアントサーバシステムのブロック構成例である。第９図において、検索キー入力手段６０１は、利用者が収集する対象を検索するために、検索のキーとなるキーワードや位置などを入力する手段である。音声送信要求手段６０２は、サーバ側に対して、音声情報の送信を要求する手段である。マルチメディア情報取得手段６０３は、送信を要求された音声情報が含まれるマルチメディア情報を図示しないデータベースから取得する手段である。音声抽出手段６０４は、マルチメディア情報に含まれる音声情報部分を抽出する手段である。音声送信手段６０５は、音声情報をクライアント側に送信する手段である。

音声検索手段６０６は音声情報について音声認識を行ない、音声認識結果に対して、検索キーとして指定された文字列について検索や話者検索を行なう手段である。なお、入力音声の音声認識を行なう方法として、例えば、入力音声を小区間ごとに音素の標準パタンと比較して距離を求め、距離の近い音素を音素認識結果として出力し、さらに音素系列を単語音声辞書と比較する手段がある（前出“ディジタル音声処理”，東海大学出版会、ｐｐ１６７－１８６　音素を単位とする単語音声認識」参照）。動画像送信要求手段６０７は、サーバ側に対して、特定の区間の動画像情報の送信を要求する手段である。シーン抽出手段６０８は、全動画像内から、指定された区間の動画像情報を抽出する手段である。動画像送信手段６０９は、クライアント側に対して、動画像情報を送信する手段である。動画像表示手段６１０は、動画像を表示する手段である。

第９図において、クライアント側において、まず、検索キー入力手段６０１を用いて入力された文字を指定文字列とする。次に、音声送信要求手段６０２において、特定のマルチメディア情報内の音声情報の送信を要求する。次に、サーバ側において、音声情報の送信要求を得たのち、マルチメディア情報取得手段６０３において、送信を要求された音声情報を含むマルチメディア情報をデータベースから取得する。さらに、取得したマルチメディア情報内の音声情報部分を、音声抽出手段６０４において抽出し、音声送信手段６０５において、音声情報部分のみをクライアントに送信する。クライアント側では、音声検索手段６０６において、受信した音声情報について、指定文字列の検索を行なう。なお、ここでは、受信した音声情報について一度音声認識を行ない、認識結果に対して指定文字列の検索を行なう音声検索方法を仮定している。次に、指定文字列が含まれる音声区間に対応する動画像の送信を、動画像送信

要求手段６０７において要求する。さらに、サーバ側では、受信した動画像送信要求に基づき、シーン抽出手段６０８において、要求された区間の動画像を全動画像から抽出し、動画像送信手段６０９によってクライアント側に送信する。

以上の構成により、音声検索が可能なマルチメディア情報検索クライアントサーバシステムにおいて、全マルチメディア情報をサーバ側からクライアントに送

信することなく、必要な情報のみを送信することが可能となる。

第１０図は、マルチメディア情報検索クライアントサーバシステムの他のブロック構成例を示す図である。

第１０図において、クライアント側において、まず、検索キー入力手段６０１を用いて入力された話者名を指定話者名とする。次に、音声送信要求手段６０２において、特定のマルチメディア情報内の音声情報の送信を要求する。次に、サーバ側において、音声情報の送信要求を得たのち、マルチメディア情報取得手段６０３において、送信を要求された音声情報を含むマルチメディア情報をデータベースから取得する。さらに、取得したマルチメディア情報内の音声情報部分を、音声抽出手段６０４において抽出し、音声送信手段６０５において、音声情報部分のみをクライアントに送信する。クライアント側では、音声検索手段６０６において、受信した音声情報について、指定話者の検索を行なう。なお、ここでは、受信した音声情報について話者識別を行ない、識別結果に対して指定話者名の検索を行なう音声検索方法を仮定している。次に、指定話者名に対応する音声区間の動画像の送信を、動画像送信要求手段６０７において要求する。さらに、サーバ側では、受信した動画像送信要求に基づき、シーン抽出手段６０８において、要求された区間の動画像を全動画像から抽出し、動画像送信手段６０９によってクライアント側

に送信する。

以上の処理より、話者検索が可能なマルチメディア情報検索クライアントーサバシステムにおいて、全マルチメディア情報をサーバ側からクライアントに送信することなく、必要な情報のみを送信することが可能となる。

第１１図は、本発明のマルチメディア情報オーサリング方式の画面表示例である。第１１図において、映像表示領域１２０１は、ディスプレイ上の、動画像を表示する領域である。インデクス表示領域１２０２は、ディスプレス上のマルチメディア情報インデクスを表示する領域である。話者名表示領域１２０３は、各音声区間に対応する話者名を表示する領域である。話者名表示方法として、各音声区間に対して話者名を表示する方法と、話者毎に分割した上で話者名を表示す

る方法が考えられる。

第１１図において、話者名表示領域１２０３に表示された話者名を元に、利用者は、キーボードなどの文字手段を用いて、人物名を入力する。あるいは、マウスなどの位置入力手段を用いて、話者名表示領域に表示された話者を指定することにより話者名を入力する。入力された話者名に基づき、話者の音声区間の音声あるいは動画像を出力させることができる。

本発明によれば、複数の話者による音声を含む映像に対して、各話者ごとの音声に対応する音声区間の音声あるいは動画像を出力させることができる。

複数の人物画像が同一画像内に存在する場合、音声区間を指定することにより、音声区間の音声に対応する人物画像、指定音声区間の音声と同一話者の全音声区間の音声に対応する人物画像、を抽出することができる。

同様に、画像を指定することにより、各話者ごとの音声に対応する音

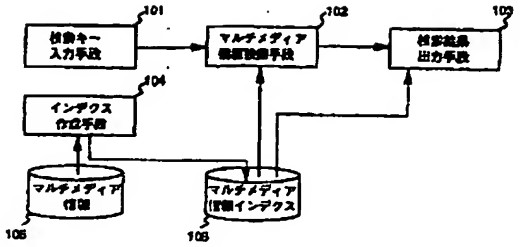
声区間の音声、動画像、あるいは、指定画像に対応する音声区間の音声と同一話者の全音声区間の音声に対応する人物画像、を出力させることができる。

産業上の利用可能性

本発明は、ＰＤＡ(Personal Digital Assistant)、ノートパソコンなどの携帯情報端末や、パーソナルコンピュータ、ワークステーションなどのマルチメディア端末当の、音響情報を含む映像を扱う機器に適する。これにより、話者別の映像を容易に抽出するオーサリング方式を備えるシステムを提供できる。

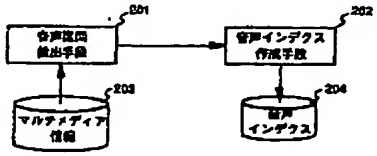
【図1】

第1図



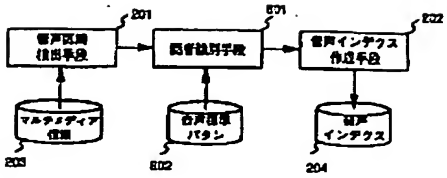
【図2】

第2図



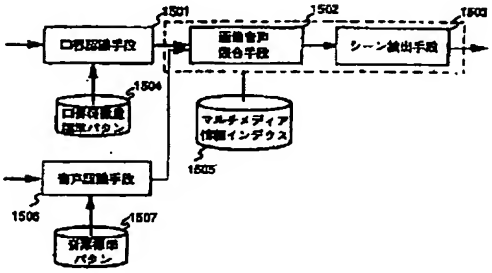
【図3】

第3図



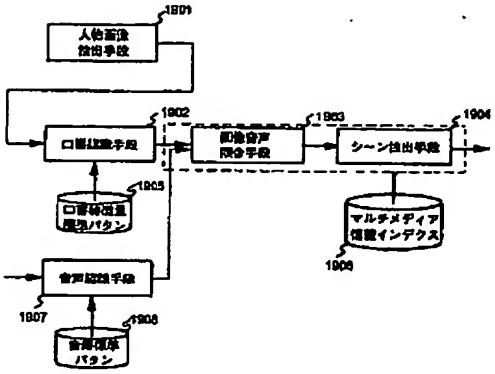
【図4】

第4図



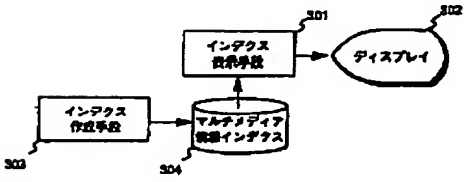
【図5】

第5図



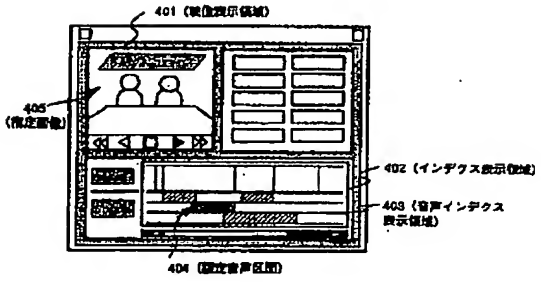
【図6】

第6図



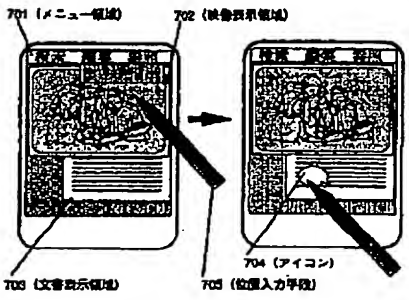
【図7】

第7図



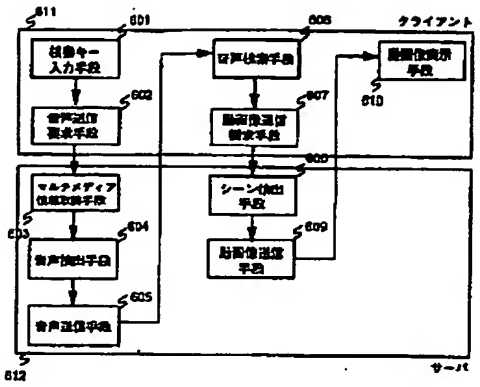
【図8】

第8図



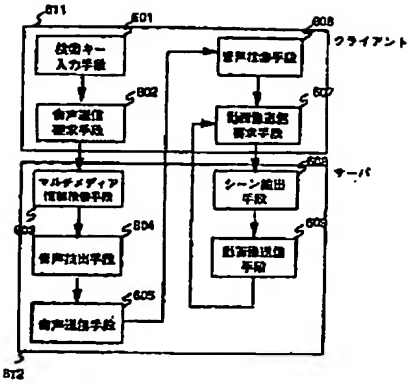
[図9]

第9図



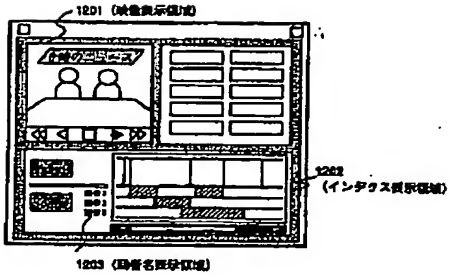
[図10]

第10図



[図11]

第11図



【国際調査報告】

国際調査報告

国際出願番号 PCT/JP 95/01746

A. 発明の属する分野の分類（国際特許分類（IPC））

Int. Cl⁶ G06F17/30

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int. Cl⁶ G06F17/30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報 1926－1994年
日本国公開実用新案公報 1971－1994年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JICST科学技術文献ファイル

C. 関連すると認められる文献

引用文献のカテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
Y	1989情報学シンポジウム講演論文集（東京），17.1月.1989（17.01.89），小川隆一他「音声，動画を含むハイパーメディア作成支援システム」第43－52頁特に50頁	1－10
Y	JP，7－226931，A（株式会社 東芝，日本電信電話株式会社），22.8月.1995（22.08.95），第1欄第2－42行（ファミリーなし）	1－10

☑ C欄の続きにも文献が列挙されている。

☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー
「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」先行文献ではあるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
17.11.95	05.12.95

名称及びあて先	特許庁審査官（権限のある職員）
日本国特許庁（ISA/JP） 郵便番号100 東京都千代田区霞が関三丁目4番3号	高原 勲 5L 9069 電話番号 03-3581-1101 内線 3564

様式PCT/ISA/210（第2ページ）（1992年7月）

国際調査報告 国際出願番号 PCT/JP 95/01746

C (続き). 関連すると認められる文献

引用文献のカテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
Y	日本機械学会東北支部・情報工学会東北支部地方講演論文集 VOL. 1993, NO. Yonezawa 中野政身他「ステレオ視による機械視座に関する研究(母音口形の識別)」第255—257, 特に255頁第1欄	8

様式PCT/ISA/210 (第2ページの続き) (1992年7月)

（注）この公表は、国際事務局（ＷＩＰＯ）により国際公開された公報を基に作成したものである。

なおこの公表に係る日本語特許出願（日本語実用新案登録出願）の国際公開の効果は、特許法第１８４条の１０第１項（実用新案法第４８条の１３第２項）により生ずるものであり、本掲載とは関係ありません。